VŠB TECHNICKÁ UNIVERZITA OSTRAVA
FAKULTA ELEKTROTECHNIKY A INFORMATIKY

Rozpoznání směru pohledu řidiče s využitím kamery ve vozidle

Gaze Recognition of Driver using In-vehicle Camera

Veronika Gromnicová

Bakalářská práce

Vedoucí práce: Ing. Michael Holuša, Ph.D.

Ostrava, 2021

Abstrakt

Tato bakalářská práce se zabývá odhadem směru pohledu řidiče s využitím metod hlubokého učení. Monitorování pohledu může snížit rizika vyplývající z nepozornosti řidiče a předcházet možným rizikům. Práce je zaměřena na existující metody pro odhad směru pohledu, jejichž implementace jsou volně dostupné. Vybrané metody jsou v práci představeny a následně otestovány na videonahrávkách snímaných během jízdy ve vozidle. Na základě tohoto testování je pak vyhodnocena úspěšnost metod, která závisí nejen na přesnosti odhadovaných vektorů pohledu, ale také na časové náročnosti, která je pro použití za provozu důležitá.

Klíčová slova

počítačové vidění, hluboké učení, detekce, směr pohledu, CNN

Abstract

This bachelor thesis deals with the estimation of driver's gaze using deep learning technology. The gaze monitoring can reduce the risks of driver's inattention and prevent potential hazards. The paper is focused on existing methods for eye gaze estimation, the implementations of which are freely available. The proposed methods are presented and tested on custom video recordings taken while driving a vehicle. Based on this testing, success of the methods is evaluated, which depends not only on the accuracy of the estimated gaze vectors, but also the time required, which is important for real-time monitoring.

Keywords

computer vision, deep learning, detection, gaze direction, CNN

Poděkování

Ráda bych na tomto místě poděkovala panu Ing. Michaelu Holušovi, PhD. za množství cenných rad a především za veškerý čas, který mi ochotně věnoval při vzniku této práce.

Obsah

Seznam použitých zkratek a symbolů

OS	–	Operační systém
NHTSA	–	Národní správa bezpečnosti silničního provozu
CNN	–	Konvoluční neuronové sítě
MTCNN	–	Víceúčelová kaskádová konvoluční neuronová síť
LSTM	–	Long short-term memory - dlouhodobá krátkodobá paměť
ROS	–	The Robot Operating System
MSE	–	Mean squared error - střední kvadratická chyba

Seznam použitých cizích termínů

open-source	–	Volně dostupný software pro veřejnost
in-the-wild	–	Zachyceno mimo laboratorní podmínky v různých prostředích
ground truth	–	Ideální očekávaná hodnota - skutečnost
real-time	–	V průběhu reálného času

Seznam obrázků

Seznam tabulek

Kapitola 1

Úvod

Tato bakalářská práce se zabývá rozpoznáváním směru pohledu řidiče na základě analýzy obrazu získaného z kamery umístěné v kabině vozidla. Moderní asistenční systémy v automobilu jsou schopny minimalizovat chybné jednání řidiče, které je častou příčinou ohrožení bezpečnosti silničního provozu. Monitorování pohledu může detekovat nepozornost či únavu řidiče a případnou signalizací snížit související rizika.

Cílem této práce je vyzkoušet a porovnat různé existující algoritmy detekce směru pohledu s důrazem na jejich úspěšnost v obrazech pořízených v kabině vozidla, kde dochází k častým změnám světelných podmínek.

Celá práce je rozdělena do samostatných kapitol, které vždy začínají seznámením s jejich obsahem. Následující kapitola se zabývá analýzou problému rozpoznávání směru pohledu, obsahuje stručné seznámení s problematikou monitorování lidského pohledu, vlastnostmi lidského oka a metodiku snímání očního pohybu.

Třetí kapitola této práce je věnována zpracování obrazu a počítačovému vidění, zabývá se digitálním obrazem a detekcí objektů v obrazech. Dále pak seznamuje s použitím strojového učení v počítačovém vidění a souvisejícími konvolučními neuronovými sítěmi.

Ve čtvrté kapitole jsou představeny vybrané metody rozpoznávání směru pohledu, jež byly v posledních letech představeny a které jsou založeny na principu hlubokého učení.

V kapitole Experiment je popsána kompletní metodika přípravy a testování vybraných algoritmů na dvou různých datových sadách a jsou zde také představeny použité detekční algoritmy. Následně je provedeno zhodnocení na vybraných metrikách, jakými jsou přesnost predikcí nebo časová náročnost výpočtu. Nalezneme zde také ukázkové snímky testovacích sad s vykreslenými predikcemi každé z metod.

Celá práce a její výsledky jsou shrnuty v závěru.

Kapitola 2

Analýza problému

Rozpoznávání směru pohledu je nejen pro člověka neodmyslitelnou součástí neverbální komunikace. Oční kontakt, orientace pohledu vůči druhé osobě, nebo přímé uhýbání pohledem jsou významnými ukazateli sociálního chování. Tyto projevy jsou tedy přirozeně častými předměty studií sociálních i psychologických věd [1].

Zájem o rozpoznávání směru pohledu se však rozvinul i v mnoha jiných odvětvích, především v těch technických. Člověk dnes například může ovládat počítač či televizi pouhým pohledem. Přední kamery chytrých mobilních telefonů mohou dle směru pohledu uživatele aktivovat funkce, jakými jsou například odemykání či zamykání. Real-time monitoring směru pohledu řidiče ve vozidle pak může být preventivním opatřením před nepozorností řidiče nebo jeho mikrospánkem [2].

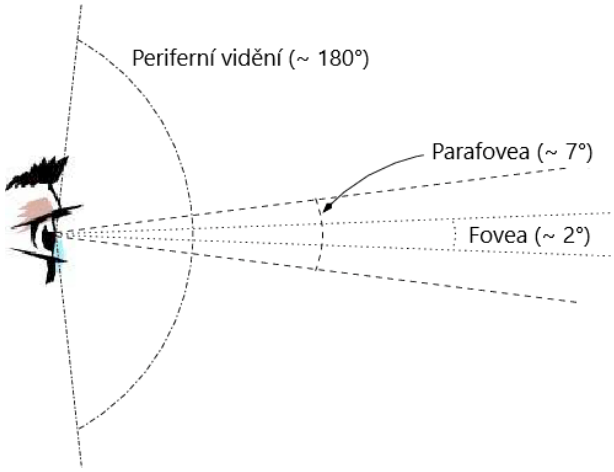
2.1 Oko

Lidské oko je pro člověka stěžejní ve vnímání a orientaci ve svém okolí. Je tvořeno vnější částí – rohovkou (cornea) a vnitřní částí – čočkou (lens). Mezi rohovkou a čočkou se nachází duhovka (iris), která reguluje množství dopadeného světla dovnitř oka. Zornicí (pupil) na duhovce prostupuje světlo do oka. Poté světlo putuje až k sítnici, na které se nachází velké množství světločivých buněk [3].

Na sítnici se nachází také žlutá skvrna (fovea), která obsahuje největší hustotu světločivých čípků. V případě, že člověk vnímá obraz přes žlutou skvrnu, nazýváme toto vidění foveálním. Naopak vnímání obrazu mimo žlutou skvrnu říkáme periferní vidění. Mezi periferním a foveálním viděním se nachází parafoveální vidění [3]. Schéma jednotlivých ostrostí vidíme na obr. 2.1.

2.2 Sledování pohybu očí

Monitorování pohybu očí je měření veškeré oční aktivity, sledování směru pohledu pak analyzuje data s ohledem na pozici hlavy i okolí sledovaného subjektu. Jak už bylo zmíněno v úvodu kapitoly, rozpoznávání směru pohledu má množství různých využití, především pro interakci mezi člověkem a počítačem, jako pomůcka pro osoby se zdravotním postižením.



Obrázek 2.1: Ostrosti vidění [4]

První studie sledování očního pohybu založená na videonahrávkách byla provedena v roce 1940 u pilotů ovládajících letadlo. Během dalších desetiletí se nadále výzkum prováděl pomocí kamer připevněných na hlavě sledované osoby. Na redukci nepohodlí subjektu a také zlepšení přesnosti se zaměřily studie po roce 1970 [5].

Směr pohledu očí může být sledován monitorovacím systémem v autě. Cílem tohoto systému je omezit počet dopravních nehod, ke kterým během jízdy dochází. Studie NHTSA z roku 2013 poukazuje na to, že nepozornost řidiče je příčinou 78 % všech dopravních nehod. Hlavní motivací k vývoji systému monitorování řidiče je tedy snížení počtu nehod a zlepšení bezpečnosti silničního provozu. Takovýto systém může detekovat například ospalost a upozornit řidiče hlasitým zvukem i vibracemi [5].

2.3 Metodika snímání očních pohybů

Způsoby, kterými se snímají oční pohyby, lze rozdělit na kontaktní a bezkontaktní (kontakt vůči oku). Bezkontaktní metody využívají především optických a elektrických vlastností oka. U kontaktních metod je třeba k oku připevnit umělé těleso, např. speciální kontaktní čočku. [6].

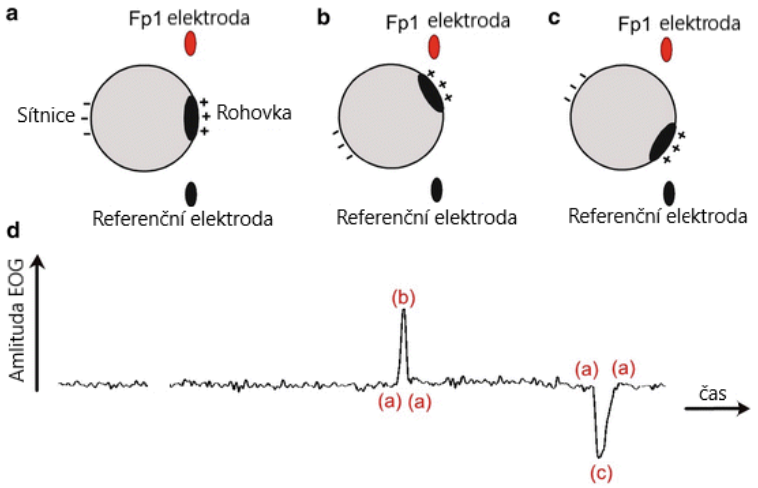
Větší část následujících metod je určena pouze pro experimentální práci, neboť jejich měření mnohdy vyžaduje laboratorní podmínky.

2.3.1 Bezkontaktní metody snímání

2.3.1.1 Elektrookulografie

Rohovka je vůči zadní části oční koule elektricky pozitivní, což vytváří potenciálový rozdíl nazývaný korneoretinální potenciál. Elektrookulografie využívá tohoto napěťového rozdílu pomocí dvou snímacích elektrod, které jsou umístěny okolo oka. Podle pohybu očí se mění vektor elektrického

pole vzhledem k elektrodám, které pak zaznamenávají horizontální i vertikální složku pohybu [7] [6], viz. obr.2.2.



Obrázek 2.2: Snímání pomocí elektrookulografie [8]

2.3.1.2 Infračervená okulografie

Při měření touto metodou se oko subjektu ozáří infračerveným světlem z nepohyblivého zdroje. Následuje analýza odraženého záření, jehož množství závisí na poloze oční koule. Infračervené záření má pouze tepelný efekt a nemá na zrak žádné rušivé účinky, nedochází ani k ovlivňování výsledků okolními světelnými zdroji [9].

2.3.1.3 Videookulografie

Při videookulografii je pohyb očí zaznamenáván kamerou na videozáznam. Ten je pak zpracováván počítačem pomocí různých algoritmů a analýzou získaných dat je následně stanoven pohyb a směr očí [10].

Pokud je videozáznam pořizován z kamery připevněné na hlavě subjektu, cílem měření je relativní poloha očí vůči hlavě. Takovéto snímání může být poskytnuto prostřednictvím speciálních brýlí, které subjekt nosí na očích po celou dobu monitorování. Takovéto brýle můžeme vidět na obr. 2.3. Pro získání absolutního směru pohledu je nutné monitorovat i polohu hlavy, k tomu lze využít právě kameru upevněnou například na palubní desce auta či v blízkosti zpětného zrcátka. Algoritmy pro rozpoznávání směru pohledu jsou dnes založeny na digitálním zpracování obrazu. Je dosahováno přesnosti na desetiny úhlového stupně a vzorkovací frekvence se pohybuje okolo stovek Hz [9].

Dříve byly využívány tzv. Purkyňovy obrázky, které vznikají při osvícení očí zdrojem světla a jsou jeho reflexním obrazem. Vytvářely se podle světlolomných ploch nacházejících se na oku [6].

Obrázek 2.3: Brýle pro snímání očních pohybů [11]

2.3.2 Kontaktní metody snímání

2.3.2.1 Magnetookulografie

Tato metoda využívá ke snímání pohybu cívku, která je složena ze závitů tenkého vodiče. Ta je připevněna k oku pomocí speciálně upravené kontaktní čočky. Je zaznamenáváno napětí, které se indukuje v cívce homogenním magnetickým polem, metoda je přesná, ale poměrně invazivní [9].

2.3.2.2 Elektroretinografie

Elektroretinografie je velmi podobná elektrookulografii, využívá tedy záznamy elektrických signálů z oka. Pro zaznamenávání signálů se taktéž využívají elektrody, jedna z nich je však připevněna přímo na kontaktní čočce na oku [12].

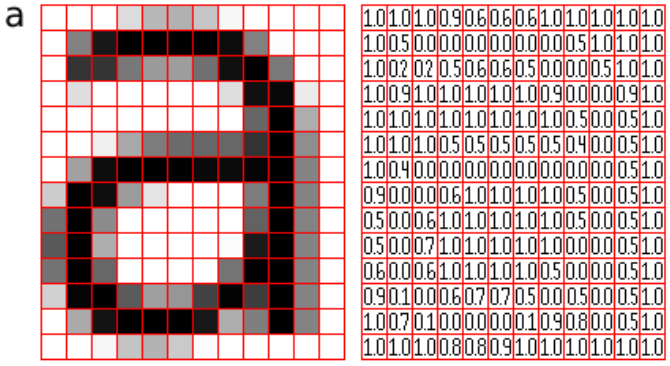
Kapitola 3

Počítačové vidění, zpracování obrazu

Tato kapitola je věnována základům počítačového vidění, které je součástí odvětví zpracování obrazu. Nejprve budou představeny základy digitálního obrazu a problematika detekování objektů a obličejových bodů. Následně bude také přiblížena role strojového učení v počítačovém vidění a související neuronové sítě.

3.1 Digitální obraz

Digitální obraz jsou strukturovaná data, která bývají obvykle reprezentována pomocí čísel (binární či hexadecimální soustavy). Obraz může být rastrový či vektorový.



Obrázek 3.1: Rastrový obraz a jeho reprezentace v číslech [13]

Rastrový obraz je popsán pomocí obrazových bodů – pixelů. Ty jsou strukturovány do matice a reprezentovány čísly (každý obrazový bod má svou hodnotu i pozici), které určují barvu jednotlivých pixelů. Hodnoty čísel mohou být například třísložkové pro barevný obraz, či jednosložkové pro obraz černobílý [3]. Rastrový obraz a jeho číselnou reprezentaci můžeme vidět na obr. 3.1.

Vektorový obraz je složen z bodů, čar a křivek, které jsou vytvořeny na základě matematických vzorců. Tvoří se tedy vektorovým popisem grafické informace. Při přiblížení vektorového obrazu zjistíme, že všechny čáry i křivky zůstávají hladké, neboť se vektorový obraz dokonale přizpůsobuje své velikosti [13] .

3.2 Detekce objektů

Počítačové vidění využívá technologie pro detekci, klasifikaci a lokalizaci objektů za účelem porozumění scénám reálného světa. Z pohledu počítačového vidění je obraz scénou složenou z objektů zájmu a z pozadí, které tvoří všechny ostatní objekty v obraze. Interakce mezi těmito objekty jsou klíčovými faktory pro pochopení dané scény. Dvěma důležitými úkoly počítačového vidění jsou detekce a rozpoznávání. Detekce objektu určuje jeho přítomnost, případně rozsah i umístění v obraze. Rozpoznávání objektu pak identifikuje třídu v trénovací databázi, do které je daný objekt zařazen. Detekce objektu tedy obvykle předchází jeho rozpoznávání a lze ji rozdělit na soft detekci, kdy je zjišťována pouze přítomnost objektu, a hard detekci, která kromě přítomnosti detekuje i umístění objektu [14].

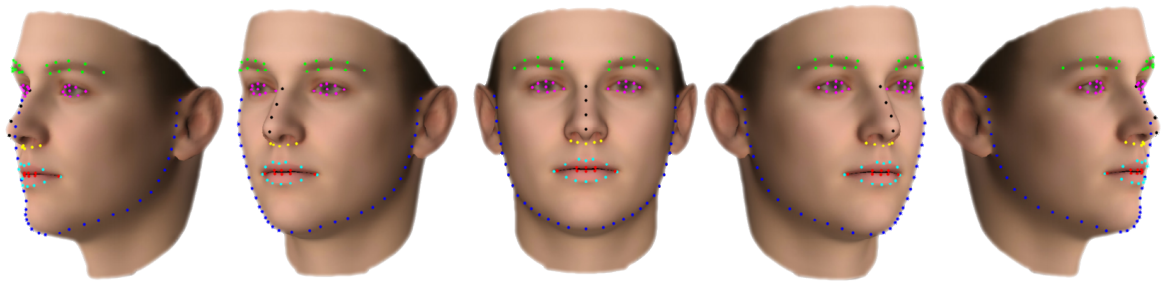
Detekce objektu se obvykle provádí prohledáním každé části obrazu, aby se lokalizovaly části, jejichž fotometrické či geometrické vlastnosti se shodují s vlastnostmi cílového objektu v trénovací databázi. Toho lze dosáhnout například skenováním šablony objektu napříč obrazem na různých místech, s různými měřítky a rotacemi. Detekce je potvrzena v případě, že je podobnost mezi šablonou a obrazem dostatečně vysoká, tuto podobnost lze měřit pomocí korelace [14].

3.3 Detekce obličejových bodů

Orientační obličejové body, facial landmarks, jsou definovány jako detekce a lokalizace určitých charakteristických bodů v obličeji. Nejběžněji používané landmarky jsou krajní body očí, špička nosu, brada nebo také krajní body úst. Je známo, že orientační body, jako je špička nosu nebo rohy očí, jsou málo ovlivňovány výrazovými pohyby obličeje, jejich detekce je tedy spolehlivá a v literatuře se označují jako výchozí (fiducial) body. Během posledních let vzrostl zájem o techniky detekce landmarků v in-the-wild obrazech [15].

Obecně můžeme landmarky rozdělit do dvou kategorií - primární a sekundární. Mezi primární orientační body řadíme ty, které hrají významnější roli v identitě obličeje. Jedná se o krajní body úst, očí, obočí a také špičku nosu, lze je relativně snadno detekovat pomocí základních obrazových funkcí. Mezi sekundární landmarky pak řadíme bradu, kontury tváří, střed obočí a úst, nosní dírky a další. Tyto body se podílí hlavně na sledování výrazu obličeje [15]. Ukázku vyznačených orientačních bodů můžeme vidět na obrázku 3.2.

Během posledních let zaznamenáváme v oblasti počítačového vidění množství výzkumů zaměřených právě na problém lokalizace orientačních obličejových bodů. Hlavním důvodem je nespočet



Obrázek 3.2: Obličejové body [16]

možných aplikací, kde hraje lokalizace obličejových rysů významnou roli, např. analýza výrazu obličeje, animace obličeje, 3D rekonstrukce hlavy, rozpoznávání tváře a jiné. Tyto aplikace pak mohou být použity pro anonymizaci identity v digitálních fotografiích, v softwaru pro úpravu tváří, odezírání ze rtů nebo interpretaci znakového jazyka [15].

Metodiku detekce landmarků v obrazech můžeme rozdělit na dva základní typy: modelový a texturový. Modelový nebo také tvarový přístup považuje soubor obličejových bodů za celistvý tvar. Metody založené na texturách hledají oproti tomu každý landmark zvlášť bez jakéhokoliv modelu [15].

3.4 Strojové učení v počítačovém vidění

Strojové učení je podoblastí umělé inteligence a jeho hlavním předmětem jsou algoritmy a techniky umožňující počítačovému systému učit se. Toto učení chápeme jako zefektivnění schopnosti adaptace na změny.

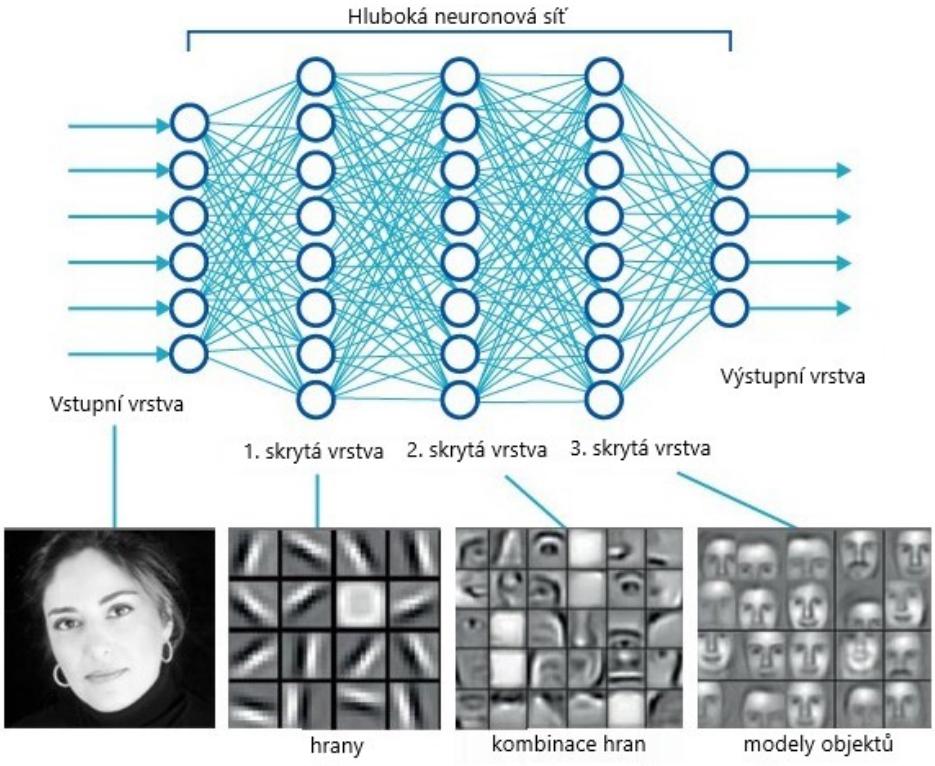
Z hlediska počítačového vidění je strojové učení schopno nabídnout efektivní metody pro automatizaci některých úkolů. Pro tento automatizační proces jsou používány tréninkové množiny vstupů, u níž jsou známy i požadované výstupy. Tyto sady vstupů poskytují vzory, dle kterých se daný systém učí a využívá je k výpočtu nových modelů [17].

Mezi podoblasti strojového učení v počítačovém vidění patří například random forests [18], algoritmy k-nejbližších sousedů [19], metody podpůrných vektorů a neuronové sítě.

Umělé neuronové sítě, ANN, jsou inspirovány nervovým systémem člověka a jsou charakterizovány, kromě schopností učení a generalizace, především možností výkonného paralelismu a tolerancí chyb i šumu. ANN využívají model funkcí biologických neuronů, tedy výkonných prvků, které jsou navzájem propojeny do sítí vazbami. Sítě mohou pracovat za pomoci vnějšího činitele - učitele, nebo na základě stimulů, kdy se jedná o tzv. samoorganizující se síť bez učitele [20] [21].

3.5 Konvoluční neuronová síť

Konvoluční neuronové sítě jsou vícevrstvé umělé neuronové sítě, které jsou úzce spjaty s metodami hlubokého učení. Úspěšnost hlubokých neuronových sítí v oblasti rozpoznávání také předčila výkony konvenčních metod zpracování obrazu.



Obrázek 3.3: Jednoduché schéma konvoluční neuronové sítě [22]

Tyto sítě se skládají z konvolučních a pooling vrstev, díky nímž jsou schopny zpracovávat velké vstupy velmi efektivně. Přitom nepotřebují takové množství parametrů jako čistá plně propojená konvoluční síť, která zpracovává vstup o velikosti totožné s rozměry vstupního obrázku [23] [24].

Konvoluční vrstva provádí operaci konvoluce. Ta má definované své konvoluční jádro, matici běžně o velikosti 3 x 3 buňky. Dekonvoluční vrstva naopak provádí inverzní operaci ke konvoluci, jejím úkolem je co nejpřesněji odhadnout vstupní matici konvoluce na základě informací o matici výstupní. Úkolem pooling vrstev je snížení výpočetní náročnosti učení sítě a zlepšení distribuce dat, čehož dosahují redukcí a opětovným zvýšením velikosti vstupních dat. Plně propojená vrstva má na starosti odpovídající spojení mezi každým vstupním a výstupním prvkem [23].

Pokud se podíváme na aktivace neuronů v jednotlivých vrstvách konvoluční sítě jako na obrázky, uvidíme, že v prvních vrstvách sítě se chovají jako detektory hran. V dalších vrstvách, které se

nacházejí v síti hlouběji, mohou neurony detekovat přítomnost stále složitějších objektů [24]. Na obrázku 3.3 nalezneme jednoduché schéma funkce CNN.

Učení neuronové sítě probíhá na základě zpětné vazby o úspěšnosti provedené predikce. Cílem je co nejvíce minimalizovat chybu a optimalizovat vnitřní parametry sítě [23]. Konvoluční síť je schopna se po adaptaci zdokonalovat bezprizorně.

Kapitola 4

Vybrané metody pro rozpoznávání směru pohledu

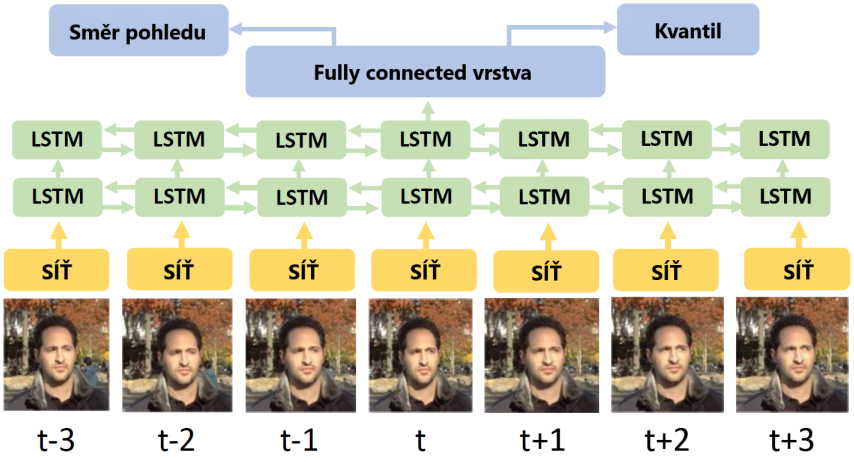
Pro tuto práci byly vybrány tři metody pro rozpoznávání směru pohledu, které využívají principů hlubokého učení. Výběr byl prováděn na základě předpokládaného využití metod při jejich vývoji. Je možné například nalézt dostupné metody, které se soustřeďují na odhad pohledu u snímků pouhých očí, což by vzhledem k účelům této práce nebylo příliš praktické. Byly tedy zvoleny metody, jenž z větší či menší části využívají k predikcím i pozici hlavy a jejichž trénovací datové sady obsahují rozsáhlejší škálu různých úhlů a pozic, které můžeme zaznamenat právě i u řidiče automobilu. Navíc byl výběr zúžen na metody, jejichž implementace je veřejně k dispozici.

4.1 Gaze360

Model metody je založen na vzhledu a využívá princip hlubokého učení. Nespoléhá se nutně na detekci očí a obličeje, což umožňuje dosáhnout vyšší robustnosti rozpoznávání v případech, kdy jsou oči například z větší části zakryty, nebo nejsou zachyceny vůbec. Metoda se tedy zabývá i případy, kdy je při natočení hlavy například viditelná pouze část jednoho oka, ale vypovídající hodnota je pro model stále dostačující. Síť poskytuje co nejpřesnější odhad i s mírou nejistoty (chybovým kvantilem) odhadu pro daný obraz. I při plně uzavřených či zakrytých očí model vydává odhadovaný směr na základě viditelných rysů hlavy, chybový kvantil se samozřejmě v takové situaci značně zvyšuje [25].

Pro co nejpřesnější výpočet metoda kromě samotného snímku využívá i okolní snímky videa, čímž se zvyšuje šance správného zachycení relevantních příznaků. Metoda ukazuje, jak využití pohybu napříč snímky pomáhá znatelně zvýšit výkon v širokém rozsahu úhlů pohledu. Na obrázku 4.1 můžeme vidět ilustraci architektury modelu Gaze360. Každý výřez obličeje je samostatně zpracován konvoluční neuronovou sítí (backbone), která produkuje vysokoúrovňové příznaky s rozměrností 256. Ty jsou dále přiváděny do dvou LSTM vrstev, které sekvence snímků štěpí do obou směrů –

dopředně i zpětně. Získané vektory jsou poté zřetězeny za vzniku dvou výstupů: predikce pohledu vůči kameře ve formátu sférických souřadnic a odhad chybového kvantilu. Jako páteřní CNN je použita ImageNet ResNet-18 [25].



Obrázek 4.1: Architektura modelu Gaze360 [25]

K předpovědi chybového kvantilu využívá metoda funkci pinball loss. Je použita jediná síť k predikci jak průměrné hodnoty, tak 10% a 90% chybového kvantilu. Je takto získán kromě odhadu pohledu také chybový kužel, kde by se ground truth hodnota měla pohybovat.

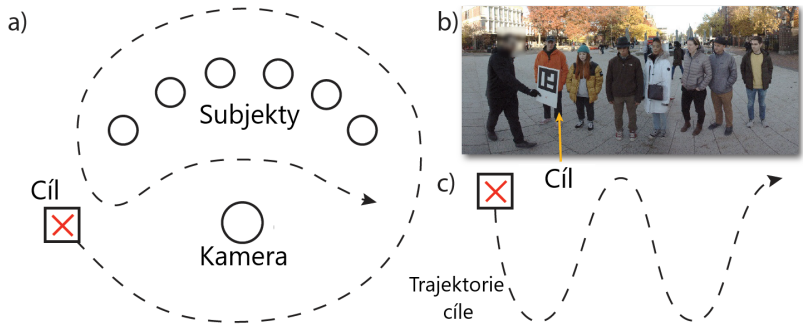
Výstupem celé sítě je $f(I) = (\theta, \phi, \sigma)$, kde (θ, ϕ) jsou sférické souřadnice odhadovaného směru pohledu, pro které je vytvořen i odpovídající ground truth vektor g v kartézském souřadnicovém systému. Vztah mezi vektorem g a sférickými souřadnicemi můžeme vyjádřit následovně $\theta = -arctan\ \frac{g_x}{g_z}$ a $\phi = arcsin\ g_y$. Poslední parametr, σ, je zmíněný offset odhadovaného směru pohledu [25].

4.1.1 Datová sada a metody jejího sběru

Cílem autorů bylo vytvořit trénovací datovou sadu, která je pořízena v nejrozmanitějších prostředích mimo laboratorní podmínky, aby byla co nejlépe aplikovatelná na in-the-wild videa. Datová sada metody Gaze360 eliminuje co nejvíce omezení pózy sledovaných subjektů a tím zajišťuje pokrytí široké škály orientací hlavy i oční bulvy vůči kameře. Snímání sady je prováděno více kamerami najednou mimo laboratorní prostředí a je pozorováno více pohybujících se subjektů zároveň. Díky pohyblivému snímání je dataset lépe přizpůsoben pro rozpoznávání v mírně rozmazaném obraze. Možné využití je proto očekáváno především ve sledovacích kamerách či interaktivní robotice [25].

Sběr datové sady byl prováděn pomocí panoramatické kamery LadyBug5 s 360 stupňovým snímáním. Tato kamera byla umístěna na stativu uprostřed scény a pozorované subjekty byly instruovány ke sledování pohyblivé desky se zobrazeným AprilTag a křížkem, na který se měly subjekty

fixovat. LadyBug5 zařízení tvoří pět synchornizovaných a překrývajících se kamerových jednotek s rozlišením 5 megapixelů a jedna kamera směřující vzhůru, která nebyla použita. Kompaktnost celého nastavení, které se skládalo z jediné kamery, notebooku a mobilního zdroje energie, umožňovala snadnou přenositelnost a tím zprostředkovala efektivní sběr dat [25]. Postup sběru datové sady nalezneme na obr. 4.2.



Obrázek 4.2: Postup sběru datové sady Gaze360 [25]

Kamera LadyBug5 poskytuje 3D paprsek v globálním kartézském systému pro každý obrazový pixel. Ten je použit k odvozování polohy nohou a očí ve sférických souřadnicích. Fixačním bodem je již zmíněný křížek na desce s AprilTag, ten je použit pro sledování desky ve 3D prostoru [25].

4.2 RT-GENE

Model metody RT-GENE je taktéž založen na principech hlubokého učení. Samotnému odhadu směru pohledu předchází několik kroků. Nejprve je použita víceúčelová kaskádová konvoluční síť (MTCNN) pro detekci obličeje a orientačních obličejových bodů - očí, nosu a úst. Využitím extrahovaných landmarků se obličej natočí a změní měřítko tak, aby byly minimalizovány vzdálenosti mezi nalezenými body a předem definovanými průměrnými pozicemi obličejových bodů. Tímto procesem vzniká normalizovaný obraz, ze kterého se následně extrahují výřezy očí. Poté je vypočítána také pozice hlavy využitím metody představené autory Patacchiola a Cangelosi [26]. Ta využívá CNN a adaptivní gradientní metody.

Vektor směru pohledu je pak prováděn pomocí navrhované sítě. Nejprve je na samostatné snímky očí aplikována síť VGG-16, která poskytuje extrakci příznaků. Každá tato síť je následována plně propojenou vrstvou o velikosti 512 po poslední vrstvě max-poolingu. Poté dochází ke zřetězení těchto vrstev a vzniká vrstva o velikosti 1024. Po této vrstvě následuje další plně propojená vrstva o velikosti 512, ke které je připojen vektor pro pozici hlavy, za kterým následují další dvě plně propojené vrstvy o velikosti 256. Výstupy poslední vrstvy jsou odhadované sférické úhly pohledu. Pro zvýšení robustnosti je použit pro celkový odhad průměr predikcí jednotlivých sítí [27].

4.2.1 Datová sada a metody jejího sběru

Trénovací datová sada metody RT-GENE byla pořizována mimo laboratorní podmínky s důrazem na větší vzdálenost subjektu od kamery. Při sběru snímků byla pro automatickou anotaci použita kombinace systému snímání pozice hlavy a mobilních brýlí pro monitoring pohledu očí. Snímací systém OptiTrack monitoruje pozici brýlí vůči kameře a navíc poskytuje RGB snímky s rozlišením 1920 x 1080 a také hloubkový snímek s rozlišením 512 x 424 pixelů.

Kompletní datový soubor obsahuje nahrávky 15 účastníků zachycených ve 122 531 trénovacích snímcích a 154 755 neoznačených snímků, kde subjekty nemají brýle [27].

Jelikož monitorovací brýle znatelně mění vzhled subjektu, bylo nutné pozdější odstranění brýlí ze snímku. Toho bylo docíleno využitím tzv. image inpaiting, konkrétně pak architektury Generative Adversial Network, kdy byla vzata v úvahu jak texturní podobnost s oblastmi v těsné blízkosti brýlí, tak sémantika obrazu. Ukázku zabarvení brýlí můžeme vidět na obrázku 4.3. Následně byly upravené snímky použity pro trénování nového modelu pro odhad pohledu.

Pro zvýšení robustnosti odhadu pohledu byla trénovací sada rozšířena hned několika způsoby. Kvůli možné chybné centralizace očí díky nedokonalé extrakci landmarků bylo nejprve provedeno 10 augmentací ořízntutím obrazu po stranách a následným zvětšením na původní velikost. Pro adaptaci na rozmazaný obraz bylo sníženo rozlišení obrazu na polovinu a čtvrtinu původního rozlišení a následnou bilineární interpolací jsou získány dva rozšiřující obrazy původní velikosti. Dále je použita ekvalizace histogramu pro pokrytí různých světelných podmínek. Nakonec jsou snímky konvertovány také do černobílých verzí [27].

Obrázek 4.3: Snímky před a po odstranění brýlí [27]

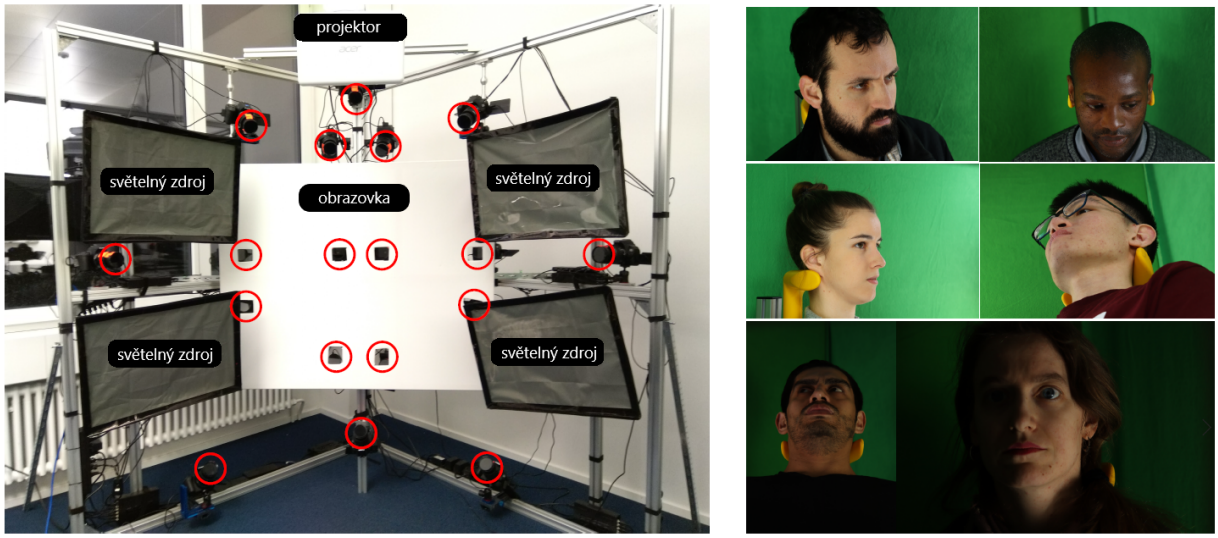
4.3 ETH-XGaze

Základem této metody je, stejně jako u předchozích metod, použití konvoluční neuronové sítě, tedy technologie hlubokého učení. Jako páteřní síť je použita běžná ResNet-50 síť. Na vstupu je očekáván obličeje o rozměrech 224 x 224 pixelů. Byl použit optimalizátor ADAM s mírou učení 0,0001 a batch size byl nastaven na hodnotu 50. Model byl natrénován pro 25 epoch a rychlost učení byla rozložena o faktor 0,1 každých 10 epoch [28].

Před aplikací modelu probíhá na snímku normalizace, po které by v ideálním případě měly oči subjektu být ve vodorovné poloze a transformované do shodné vzdálenosti od kamery. Výstupem modelu je pak směr pohledu vůči kameře po této normalizaci v podobě sférických souřadnic.

4.3.1 Datová sada a metody jejího sběru

Cílem sběru datové sady bylo především maximalizace rozsahu parametrů, které definují komplexní datový soubor pro odhad pohledu. Těmito parametry jsou především pozice hlavy, vzhled subjektu, světelné podmínky a kvalita obrazu. Velký důraz byl kladen na vysoké rozlišení snímků, což umožňuje detailnější zachycení obličejových příznaků subjektu [28].



Obrázek 4.4: Prostředí pro sběr dat (nalevo) a vzorky zachycené datové sady ETH-XGaze [28]

Datová sada byla snímána digitálními zrcadlovými fotoaparáty Canon. Pět spárovaných fotoaparátů sloužilo pro snímání geometrie a 8 kamer se postaralo o získání textury, díky které bylo možno rekonstruovat 3D tvář. Subjekt byl při snímání osvětlován čtyřmi světelnými zdroji, jejichž použitím byly simulovány různé světelné podmínky. Na velké obrazovce se před subjektem pohybovaly fixační body. Rozlišení pořízených snímků je 6000 x 4000 pixelů. Na obrázku 4.4 můžeme vidět kompletní

akvizici s 18 kamerami. Během nahrávání sedí účastníci přibližně ve vzdálenosti jednoho metru před obrazovkou, přičemž hlava je umístěna v opěrce, aby bylo zamezeno neúmyslnému pohybu hlavy. Tréninková sada je složena z 80 subjektů a obsahuje přes 1 milion záznamů [28].

Kapitola 5

Experimenty

Metody představené v předešlé kapitole byly testovány na dvou datových sadách. První datovou sadou je testovací set vytvořený autory metody Gaze360. Jejich model nebyl na dané datové sadě trénován, ale díky podobnostem mezi testovacími a trénovacími snímky byly očekávány od metody Gaze360 nejlepší výsledky. Jiné veřejně dostupné anotované datové sady pro směr pohledu obsahují snímky pouhých očí, nebo jsou na zpracování v dostupných podmínkách příliš velké, proto byla právě tato datová sada pro mou práci nejvhodnější. Druhou použitou sadou pro testování byl pak vlastní dataset v podobě videí natočených při jízdě přímo ve vozidle, kde docházelo k častým změnám světelných podmínek a snímání bylo prováděno za různých denních dob. Datová sada Gaze360 slouží spíše jako orientační přehled, neboť neobsahuje reálné snímky z automobilu, proto bude stěžejní pro hodnocení především testování na vlastních datech. Jedna z podkapitol popisuje podrobnosti obou datových množin.

Všechny výpočty při testování metod byly prováděny na osobním notebooku Dell Vostro 7500 s procesorem Intel Core i7-10750H a grafickou kartou NVIDIA GeForce GTX 1650 4GB.

5.1 Použité detekční algoritmy

Vzhledem k různým nárokům na vstupní data u každé z metod, bylo pro účely práce využito několik různých detekčních algoritmů. V následující části budou stručně představeny veškeré aplikované detekce.

5.1.1 DensePose

DensePose je algoritmus, který kromě detekce lidského těla poskytuje také segmentační zpracování. Detektor oblasti zájmu, tedy člověka, tvoří CNN s architekturou Resnet50. Detekovaný výřez tvoří vstup pro následnou část, která tvoří segmentační masku. Kompletní segmentaci obrazu provádí systém DenseReg v kombinaci s architekturou Mask-RCNN. Tato kombinace tvoří nový návrh algoritmu DensePose-RCNN [29].

Obrázek 5.1: Detekce a segmentace metody DensePose [29]

Pro naše účely není segmentační schopnost přímo využívána, ale detekce algoritmu jsou velmi spolehlivé, stejně jako následné sledování subjektu.

5.1.2 YOLO

You only look once je detekční algoritmus využívající jedné konvoluční neuronové sítě, která současně vyhodnocuje více výřezů z obrazu. Výhodou tohoto algoritmu oproti předchozím tradičním metodám je především jeho rychlost a přímá optimalizace výkonu detekce. YOLO zpracovává vstupní obrazy globálně, snímá tedy celý obraz najednou bez použití posuvného okna, což umožňuje implicitní kódování kontextu a vzhledu klasifikačních tříd [30].

Pro účely práce byl natrénován model YOLO detektoru hlavy na základě veřejně dostupné implementace [31]. Navíc byl model doplněn o implementaci jednoduchého trackingu.

5.1.3 MTCNN

Multi-Task Cascaded Convolutional Network je představen jako hluboký kaskádový víceúčelový systém k výkonné detekci obličeje. Tento rámec využívá kaskádovou strukturu tří navržených konvolučních sítí, které predikují umístění tváře a orientačních bodů způsobem hrubého přístupu. Klasifikují tedy nejprve zevrubné oblasti objektu a nepostupují tradičně od detailů objektu [32].

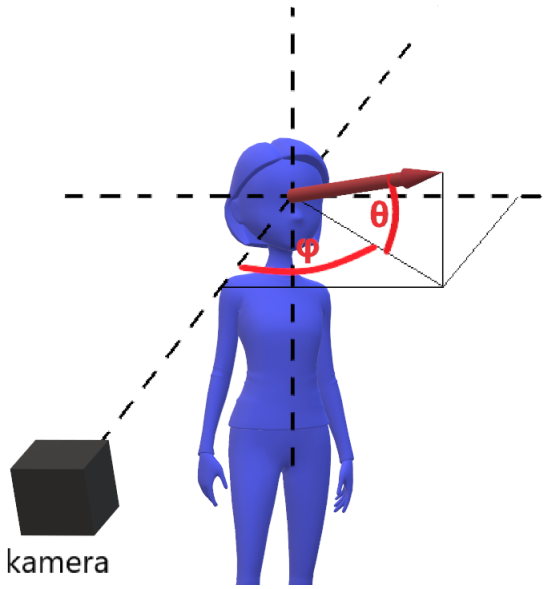
Použití MTCNN sítě bylo realizováno na základě veřejného repozitáře [33] čerpajícího z výše uvedené publikace.

5.2 Vstupní a výstupní data metod

Tato podkapitola je věnována seznámení se sférickým systémem použitým při testování metod a následně představíme pro každou metodu očekávaná vstupní data a také výstupní souřadnice algoritmů s popisem případných potřebných úprav.

5.2.1 Sférický systém pro směr pohledu

Veškerá výstupní data získána z metod byla sjednocena do sférického systému, který je znázorněn na obr. 5.2 vodorovná osa systému je shodná s náklonem snímacího zařízení a jednu z os tvoří spojnice mezi zařízením a očima subjektu. Hodnotou definující směr pohledu je tedy dvojice úhlů označovaných ϕ a θ, ty jsou v práci vyjádřeny vždy v radiánech. Ve stejném sférickém systému jsou anotovány datové sady.



Obrázek 5.2: Sférický systém pro výstupní data

5.2.2 Gaze360

Algoritmus metody Gaze360 očekává na svém vstupu výřez celé hlavy sledovaného subjektu. Pro detekci hlavy byly zvoleny dva různé detektory. Prvním z nich byl algoritmus DensePose, který poskytuje velkou přesnost, ale je poměrně výpočetně náročný. Jako druhý detekční model byl tedy zvolen detektor YOLO, který na mírný úkor přesnosti poskytuje mnohem rychlejší detekci. Pro srovnání výpočetní náročnosti bylo zachyceno, že na stejných snímcích byla detekce s použitím YOLO modelu zhruba pětkrát rychlejší než DensePose. Vzhledem k účelu této práce je výpočetní náročnost jedním z důležitých faktorů hodnocení, neboť monitorování musí probíhat v reálném čase. Pro testování na datové sadě Gaze360 byla použita pouze metoda DensePose, na vlastních sekvencích byly pak použity oba dva detekční algoritmy, aby bylo možno řádně porovnat případný rozdíl v přesnosti odhadu.

Výstupem metody jsou sférické souřadnice - úhlové výchylky vůči kameře, nebylo tedy třeba výstup metody nijak zpracovávat.

5.2.3 RT-GENE

Jak již bylo zmíněno v předešlé kapitole, detekce subjektu je již součástí metody RT-GENE. K přípravě vstupu je využita MTCNN síť, která detekuje obličej a modelu poskytuje jeho výřez. Metodě tedy na vstupu stačí poskytnout celý snímek a o jeho další zpracování se již postará sám algoritmus.

Výstupními daty metody jsou dvojice vektorů - směr pózy hlavy vůči kameře a směr natočení očí vůči hlavě. Bylo tedy třeba sjednotit tyto hodnoty pro získání celkového směru pohledu vůči kameře.

5.2.4 ETH-XGaze

Na vstupu metody je očekáván výřez obličeje subjektu. Pro tuto detekci byla zvolena stejná síť, která je použita v algoritmu RT-GENE, tedy MTCNN. Metoda byla vyzkoušena i s detektorem obličeje YOLOv2, ten však u datasetu vykazoval velké množství falešných detekcí a MTCNN byla vyhodnocena jako nejvhodnější detektor.

Výstupem algoritmu je opět odhadovaný směr ve sférických souřadnicích. Ten je ale predikován ze snímků, které prochází během výpočtu normalizací. Bylo tedy nutné výstupní vektor transformovat tak, aby odpovídal směru pohledu před normalizací obrazu. K tomu byla využita inverze rotační matice podílející se na transformaci obrazu.

5.3 Pozorované metriky

Testované metody byly vyhodnoceny na základě několika ukazatelů. Kromě přesnosti predikcí, byla věnována pozornost také časové náročnosti jednotlivých metod v kombinaci s detekcí. Dále byl vyhodnocen i úhlový rozptyl, ve kterém si metody drží určitou přesnost a porovnána byla i schopnost adaptace na zhoršení světelných podmínek a méně kvalitní obraz.

Pro určení přesnosti byla zvolena metrika MSE - střední kvadratická chyba, která je pro algoritmy regrese využívána v případech, kdy je podstatným ukazatelem velikost odchylky. Jde o veličinu statistiky vyjadřující přesnost odhadů pomocí střední hodnoty druhých mocnin rozdílů mezi odhadem a skutečností. Výpočet MSE je prováděn vzorcem 5.1 [34], kde N je počet predikcí, f je skutečná ground truth hodnota a y je metodou odhadovaná hodnota. Všechny hodnoty predikcí i odchylek jsou uváděny v radiánech.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(f_i - y_i)^2 \tag{5.1}$$

Kromě hodnoty MSE bude pro přesnost poskytnut také jednoduchý přehled poměru predikcí, které se pohybují v určité toleranci od anotovaných hodnot.

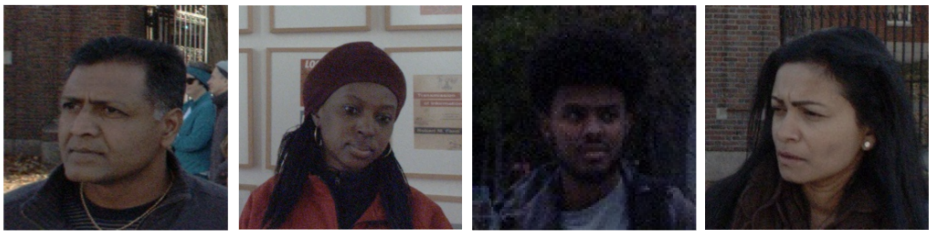
Časová náročnost jednotlivých algoritmů je uváděna v průměrném čase na jeden snímek. Naměřený čas vždy zachycuje jak výpočet predikce, tak také předcházející detekci subjektu.

Pro zhodnocení úhlového rozsahu, se kterým jsou jednotlivé metody schopny pracovat, byl zjištěn počet úspěšných odhadů u snímků s úhlem od kamery větším než 90°. Úspěšnost byla v tomto případě stanovena jako maximální odchylka 30 úhlových stupňů.

5.4 Testovací datové sady

5.4.1 Dataset Gaze360

Testovací datový set Gaze360 obsahuje 25 969 snímků, jejichž rozlišení se pohybuje mezi 106 x 106 pixelů a 300 x 300 pixelů. Jak už bylo zmíněno v kapitole vybraných metod, sběr dat probíhal na různých místech mimo laboratorní prostředí. Snímky jsou anotovány kartézskými souřadnicemi vektoru, pro vyhodnocování byly však převedeny do sférických souřadnic v radiánech. Ukázku některých snímků datasetu vidíme na obr. 5.3.



Obrázek 5.3: Ukázka snímků datové sady Gaze360

5.4.2 Vlastní dataset

Vlastní datovou sadu tvoří 17 krátkých videí, která obsahují dohromady 8526 snímků. Záznamy byly snímány v kabině vozidla za jízdy během různých denních dob - za dobrých světelných podmínek, za šera i za tmy. Některé ze snímků datové sady vidíme na obr. 5.4.

Anotace této sady byla částečně provedena využitím kombinace detekčního algoritmu DensePose a metody Gaze360, která poskytuje kromě vektoru směru pohledu také chybový kvantil odhadu. Tento kvantil byl využit pro orientační přehled o přesnosti anotace. Během testování na datové sadě Gaze360 bylo zjištěno, že pokud je chybový kvantil větší než hodnota 0,05, průměrná chybová odchylka odhadu je čtyřikrát vyšší než průměrná odchylka u predikcí s kvantilem nižším než 0,05.

Obrázek 5.4: Ukázka snímků vytvořené datové sady

Bylo tedy potvrzeno, že chybový kvantil má jistou vypovídající hodnotu. Jeho použití usnadnilo následnou manuální úpravu ground truth datové sady. Celkově bylo takto ručně upraveno přes 15 % z celkového počtu testovacích snímků, u zbylých snímků byla pouze zkontrolována přesnost stanovené anotace.

Během anotování datové sady se objevovalo množství falešných detekcí, na přiložených videích byly tyto chybné detekce ponechány a bylo nutné je ručně odstranit z anotovaných hodnot. Stejný proces byl nutný při každém využití DensePose detekce na videosekvencích.

5.5 Testování na datové sadě Gaze360

5.5.1 Přesnost

Vzhledem k tomu, že každá z metod má jiné požadavky na vstupní data, vyhodnocování přesnosti na daném datasetu nemůže být zcela jednoznačné. Sada obsahuje množství snímků, kde je subjekt otočen zcela zády a není tedy zachycen obličej. Ten je ale nutný pro predikce metod RT-GENE a ETH-XGaze, zatímco metoda Gaze360 je schopna pracovat i se snímky, které zachycují hlavu subjektu zezadu. Navíc, kvůli nutnosti použití různých detekčních algoritmů, se nepodařilo dosáhnout shodných detekcí ani u metod, které mají stejné požadavky na vstupní data. Pro každou metodu bylo tedy dosaženo jiného počtu predikcí. Metoda Gaze360 poskytla nad datasetem 24 426 predikcí (94 %), algoritmus RT-GENE detekoval pro odhad 21 892 (84 %) vstupů a u ETH-XGaze bylo dosaženo 22 016 (85 %) odhadů. Z tohoto důvodu byly hodnoty MSE v následujícím přehledu vypočítány pouze ze snímků, které byly detekovány jako vstup pro všechny tři metody. Takovýchto snímků bylo celkem 20 868.

Tabulka 5.1: MSE hodnoty metod na datové sadě Gaze360

$[rad^2]$	Gaze360	RT-GENE	ETH-XGaze
ϕ	0,292	0,498	0,521
θ	0,046	0,116	0,119

Obrázek 5.5: Ukázky zakreslených predikcí jednotlivých metod a jejich hodnoty MSE

V tabulce 5.1 si můžeme všimnout, že dle očekávání vykazuje metoda Gaze360 nejmenší hodnotu MSE. Pro úhel ϕ je tato hodnota 0,292 a pro θ bylo spočítána hodnota 0,046 rad^2. Naopak, nejméně přesnou metodou na datasetu je s hodnotou MSE 0,521 a 0,119 rad^2 algoritmus ETH-XGaze.

Můžeme pozorovat, že průměrná hodnota MSE pro úhel θ je vždy menší než pro úhel ϕ, neboť ve vertikálním směru nedochází u pohledu k takové variabilitě úhlů.

Pro lepší představu o přesnosti metod lze uvést, že do tolerance 15 úhlových stupňů algoritmus Gaze360 úspěšně odhadl směr v 82 % případů. ETH-XGaze se v toleranci udržela v necelých 61 % predikcí a RT-GENE dosáhla v této úspěšnosti přes 59 %.

Na obr. 5.5 nalezneme ukázku některých predikcí každé z testovaných metod zakreslené na snímcích datové sady a také hodnoty MSE pro tyto konkrétní případy.

5.5.2 Úhlový rozsah odhadu

Testovací datová sada obsahuje 5647 snímků, které zachycují směr pohledu od kamery v úhlu větším než 90°. Tyto snímky byly zahrnuty do předchozí statistiky přesnosti pouze v případě, že pro ně byla poskytnuta predikce od všech testovaných metod. Pro zhodnocení rozsahu úhlu, ve kterém je metoda ještě schopna odhadovat pohled, byl vyhodnocen počet úspěšných predikcí pro těchto 5647 snímků. Úspěšnost v tomto případě chápeme jako provedenou predikci a zároveň hodnotu chybové odchylky menší než 30°, tedy 0,52 radiánu.

Metoda Gaze360 predikovala úspěšně 1584 snímků, u RT-GENE algoritmu byl počet úspěšných odhadů 361 a u metody ETH-XGaze se podařilo dosáhnout úspěchu pouze u 1 snímku.

5.5.3 Časová náročnost

Na snímcích testované datové sady byl naměřen pro metodu Gaze360 průměrný čas pro jeden snímek 0,18 sekund. U RT-GENE algoritmu bylo naměřeno 0,08 sekund a ETH-XGaze metoda vyžadovala pro výpočet nad jedním snímkem průměrně 0,24 sekund.

5.6 Testování na vlastní datové sadě

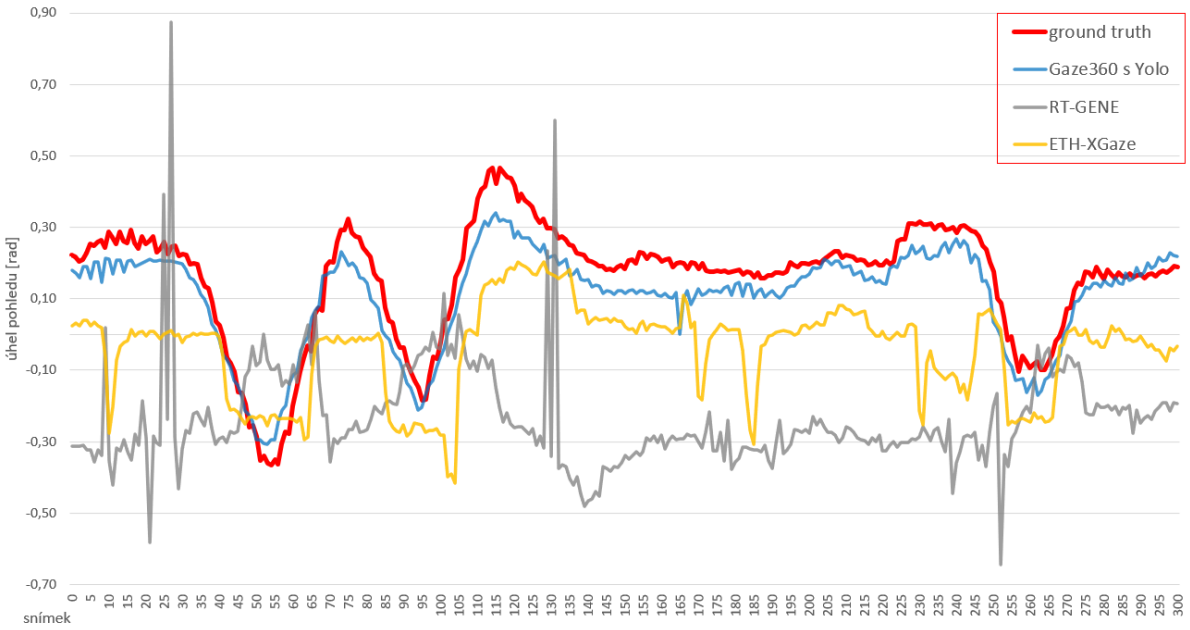
5.6.1 Přesnost

V tabulce 5.2 jsou zaznamenány hodnoty MSE pro testování na vlastních videosekvencích. Metoda Gaze360 byla otestována jak s použitím DensePose detektoru, tak s použitím YOLO algoritmu, abychom mohli zaznamenat míru ztráty přesnosti s výpočetně méně náročnou detekcí. Můžeme vidět, že zhoršení přesnosti je zanedbatelné a s použitím YOLO detekce je Gaze360 přesnější než zbylé dvě metody. MSE hodnoty s YOLO detekcí jsou pro horizontální úhel 0,008 rad^2 a pro vertikální úhel 0,003 rad^2. ETH-XGaze je druhou nejpřesnější metodou s hodnotami 0,054 a 0,03 rad^2. RT-GENE v přesnosti znatelně zaostává, jeho průměrná střední kvadratická chyba je 0,237 a 0,157 rad^2.

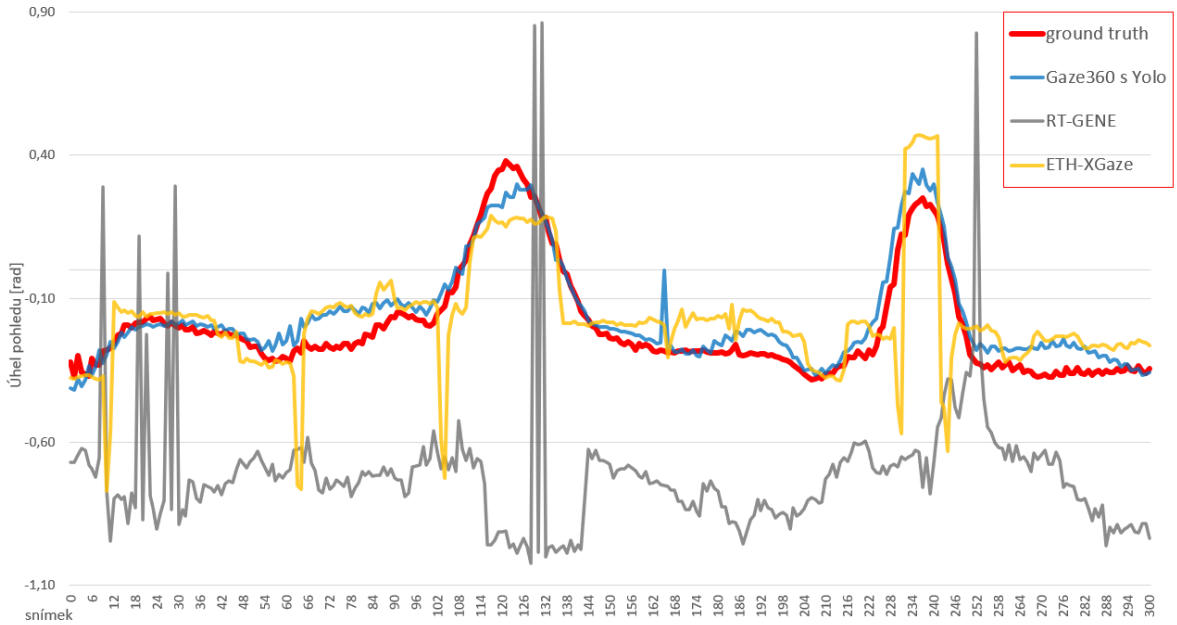
Tabulka 5.2: MSE hodnoty metod na vlastních datech

$[rad^2]$	Gaze360		RT-GENE	ETH-XGaze
	DensePose	YOLO		
ϕ	0,0008	0,008	0,237	0,054
θ	0,0003	0,003	0,157	0,03

Na grafech 5.6 a 5.7 můžeme pozorovat grafické znázornění odhadovaných úhlů oproti stanovené ground truth hodnotě.



Obrázek 5.6: Graf hodnot odhadovaných úhlů pro každou z metod - horizontální osa



Obrázek 5.7: Graf hodnot odhadovaných úhlů pro každou z metod - vertikální osa

Na obrázcích 5.8, 5.9 a 5.10 můžeme vidět ukázky predikcí metod i s příslušnými hodnotami MSE pro dané případy.

Obrázek 5.8: Ukázky vykreslené predikce metody Gaze360 s příslušnými hodnotami MSE

Obrázek 5.9: Ukázky vykreslené predikce metody RT-GENE s příslušnými hodnotami MSE

Obrázek 5.10: Ukázky vykreslené predikce metody ETH-XGaze s příslušnými hodnotami MSE

5.6.2 Přesnost za zhoršených světelných podmínek

Část datové sady byla snímána za špatných světelných podmínek - za šera i tmy. Abychom mohli zhodnotit, jak tyto podmínky ovlivňují úspěšnost metody, byly spočítány hodnoty MSE čistě pro tyto videosekvence (celkem 2332 snímků). Snímání za špatných světelných podmínek způsobuje i znatelné zhoršení kvality snímku. Hodnoty MSE nalezneme v tabulce 5.3. Můžeme vidět, že změna podmínek měla na přesnost zanedbatelný vliv a metody velmi dobře odhadují i v těchto případech. Ukázky predikce Gaze360 za zhoršených podmínek nalezneme na obrázcích 5.11.

Tabulka 5.3: MSE hodnoty metod na videích se zhoršenými světelnými podmínkami

$[rad^2]$	Gaze360		RT-GENE	ETH-XGaze
	DensePose	YOLO		
ϕ	0,002	0,018	0,246	0,035
θ	0,0004	0,004	0,113	0,026

5.6.3 Časová náročnost

Průměrný čas výpočtu nad jedním snímkem byl vždy měřen i s prováděnou detekcí. U metody Gaze360 bylo v kombinaci s algoritmem DensePose naměřen průměrný čas na snímek 0,48 s. Při použití YOLO detektoru se průměrný čas výpočtu znatelně snížil na 0,03 sekund. Metoda RT-GENE prováděla jednu predikci průměrně 0,09 sekund a u ETH-XGaze byla tato hodnota stanovena na 0,31 sekund. Metoda Gaze360 v kombinaci s YOLO detektorem byla tedy na videosekvencích nejrychlejší.

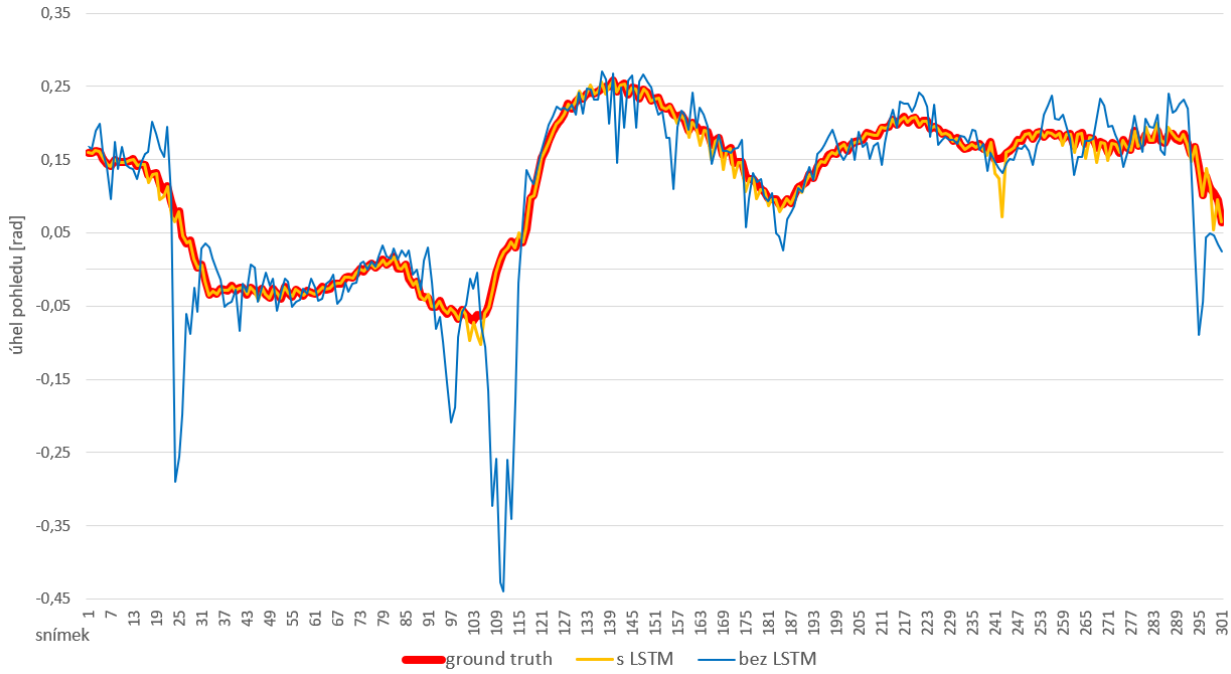
Metoda RT-GENE byla otestována i v real-time režimu s využitím ROS frameworku [35], který poskytuje prostředí pro vývoj autonomních softwarů. Výsledný stream s odhadovaným pohledem sice nebyl ve stejné rychlosti jako originální video, ale výpočet RT-GENE stíhal produkovat snímky s 20 fps. Testovací videonahrávky jsou snímány v rychlosti 30 snímků za sekundu, což není pro použití v provozu nutná hodnota.

5.6.4 Vliv použití LSTM

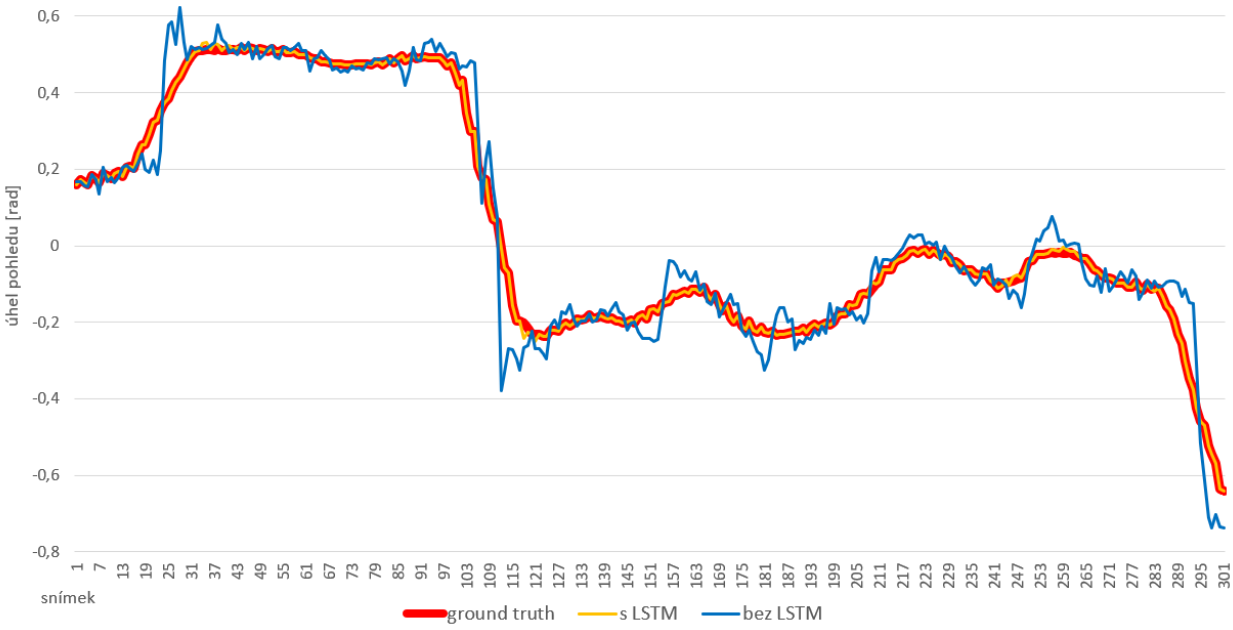
Jak již bylo dříve zmíněno, metoda Gaze360 využívá pro odhad ve videosekvencích LSTM síť, která k výpočtu odhadu pohledu těží z informací z předchozích i následujících snímků videosekvence. Tuto technologii metody RT-GENE ani ETH-XGaze nevyužívají. Pro demonstraci vlivu použití této sítě, byla otestována na všech videích metoda Gaze360 i bez využití LSTM. Pro připomenutí, MSE této metody v kombinaci s DensePose a s využitím LSTM sítě byla na hodnotách 0,0008 a 0,0003 rad^2. Při vyřazení LSTM sítě se hodnota MSE změnila u ϕ na 0,012 a u θ na 0,013 rad^2. Jde

Obrázek 5.11: Ukázky vykreslené predikce Gaze360 za špatných světelných podmínek

tedy o znatelný rozdíl, který můžeme pozorovat i na grafech 5.12 a 5.13 znázorňujících odhadované hodnoty pohledu s použitím LSTM a bez něj pro obě osy.



Obrázek 5.12: Graf hodnot odhadovaných úhlů metodou Gaze360 s LSTM a bez LSTM - vertikální osa



Obrázek 5.13: Graf hodnot odhadovaných úhlů metodou Gaze360 s LSTM a bez LSTM - horizontální osa

Kapitola 6

Závěr

Teoretickou část této bakalářské práce zahajuje stručné seznámení s problematikou monitorování očních pohybů a následně se věnuje základům a principům zpracování obrazu a počítačového vidění, které přímo či nepřímo souvisí s účelem této práce. Následně byly v textu představeny vybrané metody pro rozpoznávání směru pohledu založené na hlubokém učení, které jsou testovány v praktické části práce.

Kapitola experimenty obsahuje stručné seznámení s detekčními algoritmy, které byly při testování metod použity. Následuje představení vstupních a výstupních dat algoritmů a také pozorovaných metrik, které sehrávají hlavní roli v hodnocení testovaných metod.

Dále jsou v kapitole popsány dvě datové množiny, které byly použity pro testování metod. Jednou z testovacích datových sad je veřejně dostupný dataset Gaze360 pro detekci pohledu a druhá datová sada byla vytvořena přímo pro účely této práce. Vlastní dataset obsahuje nahrávky za reálného provozu v automobilu, kdy byly zachyceny rozsáhlé variability pohledů za různorodých světelných podmínek.

V rámci celého testovacího postupu bylo vytvořeno několik skriptů pro sběr, zpracování a vyhodnocení dat. Detailnější popis skriptů nalezneme v obsahu příloh.

Samotné testování je rozděleno v textu práce na dvě části podle testovacích dat. Nejprve jsou shrnuty výsledky získané na datové sadě Gaze360, které byly hodnoceny jak z hlediska přesnosti a rychlosti, tak z hlediska rozsahu odhadovaných úhlů. Výsledky z testování na vlastních datech jsou pro závěry práce více stěžejní, obsahují zhodnocení přesnosti za různých světelných podmínek a časové náročnosti. Úhlový rozsah není pro vlastní datovou sadu zhodnocen, neboť nahrávky obsahují zanedbatelné množství snímků zachycujících pohled ve větším úhlu než je 90°.

Na základě všech získaných výsledků můžeme jednoznačně potvrdit, že na obou datových množinách byla nejpřesnějším algoritmem metoda Gaze360. Její časová náročnost je v kombinaci s použitím Yolo detektoru také příznivá, na použité grafické kartě dosáhla rychlosti 0,03 sekund na jeden analyzovaný snímek. V kombinaci s detektorem DensePose dosahuje metoda sice o něco lepších výsledků, ale za cenu velké výpočetní a tedy i časové náročnosti, která by pro reálný provoz nebyla

použitelná. Metoda nabízí ze všech metod také největší rozsah úhlů, pro které jsou predikce poskytovány, neboť nevyžaduje nutně pro svou práci obličej. Během experimentů byl také demonstrován pozitivní vliv použité LSTM sítě.

Metoda RT-GENE byla na videonahrávkách ze všech testovaných algoritmů co do přesnosti nejméně úspěšná, vykazovala znatelné odchylky i přes poměrnou konzistenci naměřených hodnot. Ani její rychlost nepředčila předchozí metodu, na videosekvencích vyžadovala 0,09 sekund na jeden snímek.

Poslední testovaná metoda ETH-XGaze měla na datové sadě Gaze360 srovnatelnou přesnost s RT-GENE algoritmem, ale na videonahrávkách byla znatelně přesnější. Časová náročnost této metody je největší z testovaných metod, na videích se průměrný čas na jeden snímek pohyboval okolo 0,3 sekundy. Tato metoda byla autory vytvořena primárně jako základní linie pro budoucí metody rozpoznávání směru pohledu a slouží především pro základní demonstraci účelu datové sady ETH-XGaze. Dá se tedy předpokládat, že lze na tomto datasetu vytvořit propracovanější algoritmus, který by zdokonalil funkci stávajícího modelu.

Všechny testované metody vykazovaly při zhoršených světelných podmínkách pouze zanedbatelné zhoršení přesnosti.

Literatura

1. CRANACH, Mario von; ELIGRING, Johann H. Problems in the Recognition of Gaze Direction. *Introduction*. 1973, s. 419.
2. KAR, A.; CORCORAN, P. A Review and Analysis of Eye-Gaze Estimation Systems, Algorithms and Performance Evaluation Methods in Consumer Platforms. *IEEE Access*. 2017, roč. 5, s. 1. Dostupné z DOI: 10.1109/ACCESS.2017.2735633.
3. KOLEK, Petr. *Detekce směru pohledu řidiče v obrazech: Analýza problému - problematika*. Ostrava, 2017. mastersthesis. Vysoká škola báňská - Technická univerzita Ostrava.
4. NORLING, Emma. *Modelling Human Behaviour with BDI Agents*. 2009-01. Dis. pr.
5. MAVELY, A. G.; JUDITH, J. E.; SAHAL, P. A.; KURUVILLA, S. A. Eye gaze tracking based driver monitoring system. In: *2017 IEEE International Conference on Circuits and Systems (ICCS)*. 2017, s. 364. Dostupné z DOI: 10.1109/ICCS1.2017.8326022.
6. KŘÍSTEK, Jakub. *Detekce a analýza pohybu očí: Metody snímání pohybu očí*. Brno, 2013. Vysoké učení technické v Brně.
7. NEPRAŠOVÁ, Iveta. *Detekce pohybu očí: Elektrookulografie*. Brno, 2014. Vysoká škola báňská - Technická univerzita Ostrava.
8. AL-MAADEED, Somaya; BEGHDADI, Azeddine; BOURIDANE, Ahmed; CROOKES, Prof Danny; JIANG, Richard (ed.). *Biometric Security and Privacy: Opportunities & Challenges in The Big Data Era*. 1st ed. 2017. Cham: Springer International Publishing : Imprint: Springer, 2017. Signal Processing for Security Technologies. ISBN 9783319473017.
9. FEJGL, Martin. *Fokusace očí na charakteristické prvky vizuálního vjemu: Analýza očních pohybů*. Brno, 2010. Vysoké učení technické v Brně.
10. DAROFF, Robert B; AMINOFF, Michael J; DAROFF, Robert B. *Encyclopedia of the neurological sciences. Volume 1 / editors in chief Michael J. Aminoff, Robert B. Daroff. Volume 1 / editors in chief Michael J. Aminoff, Robert B. Daroff*. 2014. ISBN 9781784027988 9780123851581. OCLC: 1159664741.
11. *Tobii Pro Glasses 2* [online] [cit. 2020-12-29]. Dostupné z: https://www.tobiipro.com/product-listing/tobii-pro-glasses-2/.

12. *Elektroretinografie* [online] [cit. 2020-12-27]. Dostupné z: https://riverglennapts.com/cs/biomedical-recorders/98-electroretinography.html.

13. *Digital image representation* [online] [cit. 2020-12-31]. Dostupné z: https://pippin.gimp.org/image_processing/chap_dir.html.

14. JALLED, Fares; VORONKOV, Ilia. Object Detection using Image Processing. 2016-11.

15. OUANAN, Hamid; OUANAN, Mohammed; AKSASSE, B. Facial landmark localization: Past, present and future. In: 2016-10, s. 487–493. Dostupné z DOI: 10.1109/CIST.2016.7805097.

16. WU, Wayne; QIAN, Chen; YANG, Shuo; WANG, Quan; CAI, Yici; ZHOU, Qiang. Look at Boundary: A Boundary-Aware Face Alignment Algorithm. In: *CVPR*. 2018.

17. MAHONY, Niall O'; CAMPBELL, Sean; CARVALHO, Anderson; HARAPANAHALLI, Suman; HERNANDEZ, Gustavo Adolfo Velasco; KRPALKOVA, Lenka; RIORDAN, Daniel; WALSH, Joseph. Deep Learning vs. Traditional Computer Vision. *CoRR*. 2019, roč. abs/1910.13796. Dostupné z arXiv: 1910.13796.

18. CUTLER, Adele; CUTLER, David; STEVENS, John. Random Forests. In: 2011-01, sv. 45, s. 157–176. ISBN 978-1-4419-9325-0. Dostupné z DOI: 10.1007/978-1-4419-9326-7_5.

19. CUNNINGHAM, Padraig; DELANY, Sarah. k-Nearest neighbour classifiers. *Mult Classif Syst*. 2007-04.

20. BASHEER, Imad; HAJMEER, M.N. Artificial Neural Networks: Fundamentals, Computing, Design, and Application. *Journal of microbiological methods*. 2001-01, roč. 43. Dostupné z DOI: 10.1016/S0167-7012(00)00201-3.

21. KUČEROVÁ, Helena. *Neuronová síť* [KTD: Česká terminologická databáze knihovnictví a informační vědy (TDKIV) [online]. Praha : Národní knihovna ČR]. [B.r.] [cit. 2020-04-15]. Dostupné z: https://aleph.nkp.cz/F/?func=direct&doc_number=000000120&local_base=KTD.

22. SACHDEVA, Aashay. *Deep Learning for Computer Vision for the average person* [online]. 2017 [cit. 2021-03-25]. Dostupné z: https://medium.com/diaryofawannapreneur/deep-learning-for-computer-vision-for-the-average-person-861661d8aa61.

23. ZACHA, Jiří. *Konvoluční neuronové sítě pro klasifikaci objektů z LiDARových dat: Konvoluční neuronové sítě jejich vrstvy*. Praha, 2019. České vysoké učení technické v Praze.

24. *Neuronové sítě - konvoluční sítě a zpracování obrazu*. Dostupné také z: https://martinpilat.com/cs/prirodou-inspirovane-algoritmy/neuronove-site-konvolucni-site-zpracovani-obrazu.

25. KELLNHOFER, Petr; RECASENS, Adria; STENT, Simon; MATUSIK, Wojciech; TORRALBA, Antonio. Gaze360: Physically Unconstrained Gaze Estimation in the Wild. In: *IEEE International Conference on Computer Vision (ICCV)*. 2019-10, s. 339–357.

26. PATACCHIOLA, Massimiliano; CANGELOSI, Angelo. Head pose estimation in the wild using Convolutional Neural Networks and adaptive gradient methods. *Pattern Recognition*. 2017, roč. 71, s. 132–143. ISSN 0031-3203. Dostupné z DOI: https://doi.org/10.1016/j.patcog.2017.06.009.

27. FISCHER, Tobias; CHANG, Hyung Jin; DEMIRIS, Yiannis. RT-GENE: Real-Time Eye Gaze Estimation in Natural Environments. In: *European Conference on Computer Vision*. 2018-09, s. 339–357.

28. ZHANG, Xucong; PARK, Seonwook; BEELER, Thabo; BRADLEY, Derek; TANG, Siyu; HILLIGES, Otmar. *ETH-XGaze: A Large Scale Dataset for Gaze Estimation under Extreme Head Pose and Gaze Variation*. 2020. Dostupné z arXiv: 2007.15837 [cs.CV].

29. GÜLER, Rıza Alp; NEVEROVA, Natalia; KOKKINOS, Iasonas. Densepose: Dense human pose estimation in the wild. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, s. 7297–7306.

30. REDMON, Joseph; DIVVALA, Santosh; GIRSHICK, Ross; FARHADI, Ali. *You Only Look Once: Unified, Real-Time Object Detection*. 2016. Dostupné z arXiv: 1506.02640 [cs.CV].

31. PRANOY, Radhakrishnan. *head-detection-using-yolo* [https://github.com/pranoyr/head-detection-using-yolo]. GitHub, 2018.

32. ZHANG, K.; ZHANG, Z.; LI, Z.; QIAO, Y. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Processing Letters*. 2016, roč. 23, č. 10, s. 1499–1503. Dostupné z DOI: 10.1109/LSP.2016.2603342.

33. PAZ CENTENO, Iván de. *mtcnn* [https://github.com/ipazc/mtcnn]. GitHub, 2018.

34. *Mean Squared Error – Explained | What is Mean Square Error?* [Online] [cit. 2020-04-10]. Dostupné z: https://www.mygreatlearning.com/blog/mean-square-error-explained/.

35. STANFORD ARTIFICIAL INTELLIGENCE LABORATORY ET AL. *Robotic Operating System*. 2018-05-23. ROS Melodic Morenia. Dostupné také z: https://www.ros.org.

Příloha A

Archiv

Příloha v IS EDISON.

A.1 Obsah archivu

- dílčí implementace testovaných metod
- získaná data pro vyhodnocování
- skripty pro zpracování a vyhodnocování získaných dat
- ukázkové snímky vykreslené predikce
- část vytvořené testovací datové sady a odkaz na stažení kompletní sady
- ReadMe textový soubor s detailnějším popisem obsahu příloh